BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: IM.Sc. **SEMESTER : VIII**
BRANCH: QEDS **SESSION : SP/2025**

SUBJECT: ED417 ALGORITHMS FOR BIG DATA - I

TIME: 3 Hours **FULL MARKS: 50**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

			CO	BL
Q.1(a)	Describe the concept of Salton's Vector Space Model (VSM). Use the concept to find which two documents are most similar. You have the following four documents: Doc1: "deep learning models outperform classical methods" Doc2: "machine learning is a subset of artificial intelligence" Doc3: "deep neural networks require a lot of data" Doc4: "classical algorithms can be effective on small datasets"	[5]	1	1,2
Q.1(b)	Movie ratings of three people for four movies are given as $\begin{array}{r} \\ \text{Star Wars} \\ \text{Blade Runner} \\ \text{Amelie} \\ \text{Delicatessen} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Ali} & \text{Beatrix} & \text{Chandra} \end{array} \\ \begin{bmatrix} 5 & 4 & 1 \\ 5 & 5 & 0 \\ 0 & 0 & 5 \\ 1 & 0 & 4 \end{bmatrix} \end{array}$ The assumption are 1. All viewers rate movies consistently using the same linear mapping. 2. There are no errors or noise in the ratings. 3. We interpret the left-singular vectors u_i as stereotypical movies and the right-singular vectors v_j as stereotypical viewers. Evaluate Singular Value Decoposition for the given ratings and anlyze the same in detail. For finding Structure in Movie Ratings and Consumers, which methodology will be best suited.	[5]	1	2,3
Q.2(a)	Given a stream of elements: S={a,b,c,a,b,d,e,f,g,h,i,j}, Estimate the number of distinct elements using the Flajolet-Martin algorithm. The Hash function could be used as per your choice.	[5]	2	2,3
Q.2(b)	A search engine receives 10 million search queries per day. Each query is a string, but you don't have enough memory to store all of them. You want to find all queries that appear more than 1% of the time. Set up Misra-Gries algorithm with k= 100 counters and process the stream of queries. After processing, how would you list candidates for the "frequent queries"? Describe, why none of the actual frequent items are missed	[5]	2	2,3
Q.3(a)	Imagine you're tasked with modeling a social network of 500 users. Each user has a 5% chance of interacting with another user. This is modeled as G(500,0.05) Tasks: • Phase Transition Analysis: What is the expected number of connected components as p increases from 0.01 to 0.10 in steps of 0.01? • Giant Component Threshold: At which value of p do you start observing a giant component that contains more than 50% of the users? • Connected Network: What is the minimum p required to make the graph connected with high probability?	[5]	3	3,4

PTO

Q.3(b) Suppose you are managing a small social network with 10 users: [5] 3 3,4

They are connected like this (undirected edges):

Edge | Weight

(1,2) | 1
(1,3) | 1
(2,3) | 1
(3,4) | 2
(4,5) | 2
(5,6) | 2
(6,7) | 1
(7,8) | 1
(8,9) | 1
(9,10) | 1
(1,10) | 1

The graph is a bit dense (11 edges for 10 nodes). Use the Speilman-Srivastava Sampling idea to create a sparser graph (~4-5 edges) that preserves connectivity behavior reasonably.

Q.4(a) A retail store wants to segment its customers based on their purchasing behavior. [5] 4 4

You collected the following data from 6 customers:

Customer	Annual Spend on Electronics ($)	Annual Spend on Clothing ($)
A	1000	200
B	950	220
C	250	1200
D	300	1100
E	900	250
F	280	1000

Use Fuzzy C-Means Clustering to assign customers to two soft clusters:

- Electronics-heavy shoppers
- Clothing-heavy shoppers

Q.4(b) Suppose you are building emergency service centers in a small city. [5] 4 4

The city map has 10 neighborhoods located at coordinates:

Neighborhood	(x,y)
A	(1,1)
B	(2,3)
C	(5,4)
D	(7,2)
E	(8,5)
F	(9,1)
G	(3,7)
H	(6,6)
I	(7,8)
J	(9,9)

Choose k=3 centers such that the maximum distance from any neighborhood to its nearest center is minimized.

Q.5(a) Imagine you work for a news aggregator company that collects a lot of articles daily from various sources. You want to automatically categorize these articles into several topics without any labeled data. [5] 5 4,5

The company receives 6 articles on the following topics:

Article	Text (Summary)
1	"Stock market falls after global economic slowdown"
2	"Apple announces new iPhone and releases new features"
3	"Tech stocks rise with market optimism"
4	"New AI models outperform in image recognition tasks"
5	"Researchers develop a new algorithm for machine learning"
6	"Scientists discover new methods for climate change mitigation"

Use Latent Dirichlet Allocation (LDA) to extract 2 topics from these articles. Topic 1 should relate to finance, and Topic 2 should relate to technology/AI.

Q.5(b) Suppose you have 1000 points in a 5000-dimensional space. You want to project these points into a lower-dimensional space while approximately preserving all pairwise distances (with distortion no more than 10% — that is, distances should stay within 90% to 110% of their original values). [5] 5 5,6

What is the minimum number of dimensions you should project into, according to the Johnson–Lindenstrauss Lemma? If you use a random Gaussian projection, what properties should your projection matrix have?

:::::28/04/2025 E:::::